FÉVRIER 2022

LIVRE BLANC

LES CONCEPTS-CODE : LA SOLUTION CONTRE LES CYBERATTAQUES ?

GLIMPS

TABLE DES MATIÈRES

Cyberattaques : des tendances nouvelles, un écosystème malveillant qui se structure

Plus fréquentes depuis le début de la pandémie, les cyberattaques répondent désormais à des tendances nouvelles. Elles connaissent également une structuration des cyberattaquants, avec des logiques de division des tâches qui rendent les attaques plus dangereuses.

L'écosystème de la cybercriminalité semble avoir opéré ces derniers mois une mue. Identifiée comme étant accélérateur de tendance, la pandémie du Covid-19 a en cela provoqué une nette progression des cyberattaques. Selon l'Agence Nationale de Sécurité des Systèmes d'Information (ANSSI), entre 2019 et 2021, les signalements d'attaques par rançongiciels ont augmenté de 255%. Et tout porte à croire que cette accélération est appelée à perdurer.

Deux évolutions qualitatives sont actuellement repérées par les experts : la première a trait à la nature des attaques qui se produisent, et tout particulièrement aux tendances nouvelles qui se dessinent. La seconde est propre à l'organisation de l'écosystème cyber lui-même.

Attaques aux rançongiciels aux Etats-Unis : chiffres clé

- **30%** : c'est l'augmentation des « SAR » (Suspicious Activity Report), recensés aux Etats-Unis entre 2020 et 2021 par l'ensemble des institutions financières américaines pour le compte du FinCEN, l'agence fédérale américaine de lutte contre la criminalité financière en ligne [FinCEN, 2021].
- **590 millions de dollars** : c'est la valeur totale de l'activité suspecte liée aux ransomwares calculée sur le premier semestre 2021 aux Etats-Unis. A titre de comparaison, ce chiffre s'élevait à 416 millions de dollars pour la seule année 2020…
- **283%** : c'est l'augmentation de l'activité suspecte liée aux ransomwares entre 2020 et 2021, aux Etats-Unis.
- **5,2 billions de dollars** : c'est le chiffre total généré par les ransomwares aux Etats-Unis sur 10 ans (2011-2021).

Trois tendances nouvelles

Big Game Hunting, Raas, double extorsion… L'année 2021 aura été marquée par l'essor de trois tendances qui se combinent les unes avec les autres.

Apparu en 2018, le *Big Game Hunting* (« chasse au gros gibier ») mobilise des méthodes et des techniques d'attaques qui étaient autrefois l'apanage des opérations d'espionnage informatique initiées par des attaquants étatiques. Depuis plusieurs mois, ce type d'approche gagne du terrain et a pour objectif d'affaisser l'organisation publique ou privée dans son ensemble. Dans ce contexte, les entreprises sont ciblées en amont, et les attaques minutieusement préparées. Les données se trouvent soudainement exfiltrées avant chiffrement, et une demande de rançon est adressée au responsable de l'organisation.

S'il ne paie pas, celle-ci est mise en péril [Securelist, 2020 ; Packetlabs, 2021].

Seconde tendance : *le Ransomware as a service* (Raas). Mis en place par les cybercriminels à l'attention d'autres cyberattaquants, celui-ci repose sur le principe du Saas, avec un abonnement classique incluant tout ce dont un cybercriminel a besoin pour lancer une attaque par ransomware. Ce modèle a pour effet de démultiplier les attaques par rançongiciels tout comme les méthodes adoptées dans ces entreprises de compromission. Ces Raas se caractérisent par ailleurs par leurs mutations permanentes [CERT, 2021].

Troisième tendance : la double extorsion. Apparue en 2019, elle repose non seulement sur l'extorsion des données d'une entreprise mais également sur la menace de les publier sur un site Internet, voire auprès des médias [Threatpost, 2020 : Twitter, 2021].

Un écosystème désormais structuré

Ces pratiques témoignent d'une structuration progressive de l'univers des cyberattaquants eux-mêmes. Car l'évolution des techniques décrite précédemment est désormais l'apanage de groupes organisés, au sein desquels la division des tâches est en quelque sorte institutionnalisée. Vente de codes malveillants, mise à disposition de données personnelles, catalogues proposant des accès compromis... L'écosystème de la cybercriminalité dispose de plus en plus de ses propres produits, de ses acheteurs et de ses vendeurs identifiés [Institut Montaigne, 2021]. On y trouve désormais des services accessibles en ligne, à l'image de la location d'infrastructures de déni de service ou de processus d'anonymisation. Les cybercriminels sont en capacité de sous-traiter certaines actions, comme par exemple la compromission du SI de leurs cibles adossée à des *botnets* de « pourriels ». Ces services de distribution infectent des cibles via des mails d'hameçonnage, disséminent au cœur des SI des codes malveillants, ouvrent à leurs clients l'accès au SI compromis. Le marché du Dark Web permet l'embauche de cyberattaquants, voire de proposer des missions de courte durée ciblées et

techniquement définies.

Ces attaques structurées et de plus en plus fréquentes génèrent d'importants profits, en Europe mais également aux Etats-Unis où l'activité suspecte liée aux ransomwares a permis d'accumuler entre 2020 et 2021 un chiffre d'affaires de quelque 590 millions de dollars (voir encadré). Face à cette sophistication, l'un des enjeux est technologique, et vise à mobiliser l'Intelligence artificielle. Comme le rappelle Olivier Gesny dans la *Revue de Défense nationale*, l'IA doit contribuer à réduire le risque cyber, notamment en améliorant les capacités cognitives [Gesny, 2019].
C'est précisément ce que fait le concept-code.

Santé, éducation, ESN, collectivités locales : des secteurs ciblés

Si les cyberattaques concernent globalement toutes les zones géographiques et peuvent potentiellement frapper chaque personne et organisation, certains secteurs sont plus concernés que d'autres :
• **Santé** : les hôpitaux comme les structures de santé constituent une cible pour les cyberattaquants, en lien avec le contexte pandémique du Covid-19 [Bloomberg, 2020].
• **Education** : aux Etats-Unis, ce secteur est le second le plus attaqué après les collectivités locales [Enisa, 2020]. En France, il est moins sujet aux cyberattaques [CERT, 2021].
• **Entreprises de services numériques (ESN)** : ce secteur est particulièrement visé par les cyberattaquants.
• **Collectivités locales** : communes, intercommunalités, départements et régions sont, en France comme dans le monde, la cible des opérateurs de rançongiciels. Les mairies sont notamment visées, en raison du faible niveau de sécurité de leurs SI, de la présence de données sensibles et d'une rupture d'activité problématique.

Les cyberattaques en quelques chiffres

L'estimation des données – notamment financières – inhérentes aux cyberattaques est difficile à effectuer tant les sources sont multiples (les paiements de rançons sont souvent discrets). Ci-après, un recensement des principales données disponibles auprès des institutions spécialisées :
• **2240 attaques** par jour ont été recensées en 2021. En moyenne, une attaque est lancée toutes les 39 secondes.
• Les dépenses mondiales en cybersécurité ont dépassé les **1000 milliards de $ en 2021.**
• Les coûts des dommages liés à la cybercriminalité ont atteint **6000 milliards de $ en 2021.**
• **9 Français sur 10** ont déjà eu à faire face à un acte de cyber-malveillance.
• En 2022, **6 milliards de personnes** seront susceptibles de subir une cyberattaque.

Les campagnes de cyberattaques révèlent d'importants retours sur investissement. Selon une analyse menée en 2021 par l'Institut Montaigne et reposant sur des éléments du CERT-Wavestone, le gain net d'une cyberattaque menée selon le mode Raas (*Ransomware as a service*) serait compris entre 500K$ et 1,5M$. Soit **un ROI (retour sur investissement) compris entre 232 et 880%.**

Sources : CERT, baromètre CESIN, Cyber'Occ, Cybermalveillance, Institut Montaigne, Kaspersky Lab, Ministère de l'Intérieur, Wavestone

Le concept-code : une technologie révolutionnaire

Inhérent aux applications comme aux virus, le code informatique est au cœur de la nouvelle lutte contre les malwares. Face à ces derniers et grâce à l'IA (Intelligence Artificielle), le concept-code constitue une technologie de rupture qui se concentre sur l'histoire portée par le virus et non plus sur les mots et la grammaire auxquels il est adossé. Explications ci-après.

En 2022, le nombre d'applications téléchargées à l'échelle mondiale devrait atteindre 258 milliards. Cette dynamique ne cesse de prendre de l'ampleur : en 2017, 178 milliards d'applications faisaient l'objet de téléchargements – soit une augmentation de 45% en l'espace de cinq ans. En plus d'obéir à une tendance structurelle, cette dynamique se trouve renforcée par le contexte de crise sanitaire. En l'espace d'une année, entre 2020 et 2021, le marché a réalisé une croissance attendue initialement sur 2 à 3 ans...

Une difficulté croissante à identifier les malwares

Cette profusion en appelle mécaniquement une autre, liée à l'essor des malwares. Déjà majeures, les attaques par rançongiciels adressées à des organisations publiques et privées ont nettement cru : entre 2019 et 2020, l'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI) a enregistré une hausse de 255% [*à lire en pages précédentes*]. Pour chaque virus résident, un code viral s'installe dès lors qu'un utilisateur démarre une application logicielle contaminée.

IA, Machine Learning et Deep Learning

Le moteur d'IA qui est développé par Frédéric Grelot et ses équipes d'ingénieurs repose sur deux éléments : un algorithme qui sait lire du code et une base de connaissances qui s'enrichit sans cesse. « *Nous entraînons quotidiennement un algorithme de Deep Learning à reconnaître n'importe quelle histoire racontée dans un virus informatique* », explique l'ingénieur, ancien cadre de la DGA et co-inventeur du concept-code avec Cyrille Vignon. « *Cet algorithme reproduit en quelque sorte l'action du cerveau humain au moment de l'apprentissage de la lecture : on l'amène à travailler sur des concepts, à les reconnaître, à les comparer. Au cours de l'année 2020, nous avons ainsi effectué plus de 130 000 entraînements. À chaque campagne d'entraînements, nous confrontons nos milliards de codes informatiques à notre intelligence artificielle et lui faisons effectuer 4 mois d'entraînements – auxquels nous ajoutons des séances de réentraînement pendant 2 à 3 semaines. Nous lui faisons lire une histoire afin que l'IA crée sa propre base de connaissances. Tous les concepts de codes malveillants doivent à terme passer à ce tamis.* » À ce stade, la technologie permet de repérer 100 millions de concepts-code, ainsi que des virus actifs depuis 10 ans.

Le virus demeure même lorsque l'utilisateur a quitté l'application, constituant un agent aussi actif qu'il est discret.
À chaque application, à chaque site Internet, à chaque jeu vidéo son codage, c'est-à-dire son langage de programmation. Les virus ne font pas exception à la règle, à ceci près que le code qu'ils développent – la narration en quelque sorte – est malveillant. Ce code demeure essentiel afin de se propager sur d'autres ordinateurs, tablettes et smartphones.
Et il est de plus en plus difficile de repérer les charges malveillantes dissimulées au sein des applications, notamment en raison du recours de plus en plus fréquent à l'Intelligence Artificielle par les attaquants eux-mêmes. Les cybercriminels mobilisent en effet de plus en plus l'IA lorsqu'ils créent leurs malwares. Plus ou moins modifiés d'une version à une autre, ils présentent une signature légèrement différente qui crée des variants. Ce qui leur permet d'échapper à la vigilance des antivirus classiques [*lire encadré*]...

La conceptualisation du code : une technologie différenciante

Face à ces attaques de plus en plus sophistiquées, les lignes de défense se resserrent. Depuis quelques mois, une technologie nouvelle permet de démasquer ces malwares. Son nom : le concept-code, ou conceptualisation du code. La particularité de cette technologie qui fonctionne selon une approche articulée de *Reverse Engineering*, de *Deep Learning* et *Machine Learning* [*lire encadré*] est sa capacité à identifier des malwares jusqu'alors jamais repérés. « *Nous sommes sur un marqueur différenciant par rapport à d'autres technologies* », explique Frédéric Grelot, l'un des ingénieurs à la base de cette invention. « *Le code répond en effet à des fonctionnalités qu'il convient de percer.* »

Comprendre le concept-code grâce aux... contes pour enfants !

L'analogie entre le concept-code et les contes que l'on raconte aux enfants permet de mieux saisir les caractéristiques de cette technologie nouvelle. Car si les contes pour la jeunesse racontent une même histoire (de Boucle d'Or au Petit Chaperon rouge), celle-ci se décline auprès des enfants dans toutes les cultures et dans de nombreuses langues. C'est à ce niveau que se situe la spécificité du concept-code : il se focalise sur l'histoire contée et non sur les mots ni la grammaire employés. Ce faisant, cette technologie permet de repérer un scénario identique à 80, 90 ou 95%, dans plusieurs malwares qui pourtant sont rédigés de manières différentes. C'est donc bien l'histoire qui est au cœur du concept-code... mais aussi à la base des émotions des enfants face aux contes et légendes.

Qu'est-ce que le concept-code et comment cette technologie révolutionnaire fonctionne-t-elle ? Afin de le comprendre, il convient de revenir à la source, c'est-à-dire au code lui-même. Python, PHP, C, C++, Javascript, Java... Chacun de ces langages de programmation constitue en quelque sorte une langue différente, qui utilise sa grammaire propre. « *C'est en effet comme une langue étrangère* », explique Frédéric Grelot. « *Grâce aux lignes de code, un virus nous raconte une histoire : c'est une mauvaise histoire mais c'est une histoire tout de même. Pour cela, le malware utilise le langage du code, qui est en fait comme les mots et la grammaire d'un livre. Ce qu'il faut comprendre c'est que la technologie du concept-code se concentre sur l'histoire qui nous est racontée bien plus que sur les mots employés, car souvent les virus nous disent la même chose avec des mots différents.* » Cette nouveauté a son importance lorsque l'on sait que les antivirus actuels se concentrent davantage sur les « mots » que sur l' « histoire » qui se trouve développée... Grâce à elle, le malware ne passe plus les barrières de sécurité avec autant d'aisance qu'auparavant. « *Notre marqueur est différent des autres technologies, et c'est ce qui rend le concept-code unique* », conclut Frédéric Grelot.

Désassembler le code : l'apport du *Reverse Engineering*

La conceptualisation du code repose sur le *Reverse Engineering* automatique. Celui-ci permet de désassembler le code, c'est-à-dire de le traduire en langage compréhensible, avant d'en identifier les concepts. Ce processus se fait en plusieurs étapes, qui sont similaires à l'apprentissage de la lecture pour tout être humain : prise de connaissance des lettres d'un alphabet afin de former des phrases, puis apprentissage de la grammaire. Le point ultime de l'apprentissage étant l'assimilation de l'histoire qui se trouve narrée, avec ses concepts et ses différentes idées.

10 millions de malwares recensés : une étape vers la Cyber Threat Intelligence

La structuration de la lutte contre les malwares et les cyberattaquants passe par une importante action d'automatisation dans l'analyse des binaires. Comment celle-ci est-elle actuellement réalisée et de quelle manière le concept-code joue-t-il un rôle central dans ces actions ? Retour technique sur les différentes phases de collecte des données et leurs applications concrètes pour repérer les cyberattaquants.[1]

Lorsqu'une attaque de malware survient, plusieurs questions centrales se posent en termes de *Cyber Threat Intelligence* : Qui est l'attaquant ? De quelle manière agit-il ? À ce stade, il est primordial de saisir la nature intrinsèque de l'attaque, mais aussi et surtout d'en capter les signaux faibles. Le temps est alors précieux pour que le système de défense puisse se déployer afin de limiter, voire de circonscrire, les dommages. Dans ce contexte, il est essentiel de capitaliser, en amont, un ensemble de données sur les malwares et leurs familles d'appartenance. Une automatisation désormais possible, que la technologie du concept-code contribue à améliorer.

Collecter les données

Réalisée à grande échelle, l'automatisation de la collecte de binaires permet d'effectuer une analyse du malware de bout en bout, et ce en un temps record. Cette approche passe par plusieurs étapes clés, qui vont de la collecte des malwares à leur traitement dans un pipeline qui aboutit à la transformation en concept-code.

L'inventaire détaillé des malwares s'effectue principalement via des sources ouvertes de binaires. Dépôts Linux, dépôts Opensource type Github, binaires propriétaires, plateformes telles que Conan.io, Chocolatey, Malware-Bazaar, VX-Underground ou VirusShare… Les sites de référence sont à la fois nombreux et variés. Certains recensent des logiciels non malveillants (« goodwares »), d'autres des malwares, en proposant parfois leurs codes sources. Le nombre de fichiers accessibles est varié, allant de 100 000 fichiers à plus de 10 millions. Dans la plupart des cas, les virus recensés ont moins de trois ans. Certains sites proposent toutefois une traçabilité des malwares sur dix ans ou plus, avec des facilités d'accès plus ou moins prononcées. Le pipeline de collecte des données est décomposé en trois étapes : l'ingestion, le *tagging* et enfin la transformation en concept-code.

[1] *Cette page a été réalisée à partir des éléments de l'article scientifique suivant : GRELOT Frédéric, LARINIER Sébastien et SALMON Marie, « Automatisation de l'analyse des binaires : de la collecte source ouvert à la Threat Intelligence »,*

Ingérer les données

La première phase d'ingestion se subdivise elle-même en plusieurs sous-étapes : téléchargement, extraction, filtrage et stockage des données. Le téléchargement s'effectue par *wep scrapping*, via la plateforme Beautiful Soup. Il permet de rechercher des liens à télécharger au sein des pages des différentes sources de données.
La phase d'extraction est spécifique : elle dépend en effet de chaque source de données, et permet d'extraire des informations supplémentaires qui dépendent de la source. À titre d'exemple, des fichiers téléchargés depuis la plateforme de sources VX-Underground permettent d'identifier chaque type de malware, mais également la famille d'attaquants à laquelle il est rattaché (APTxx). Ces informations seront exploitées dans la phase de *tagging*.
La phase suivante de filtrage des données poursuit deux objectifs. D'une part, il s'agit de calculer la ressemblance du malware identifié avec d'autres malwares se trouvant dans la base. Cela est rendu possible en s'appuyant sur le hash SSDEEP. Un second hash (le SHA256) permet de s'assurer que le téléchargement en cours n'est pas celui d'une archive

Cas d'usage : liens avérés entre les familles Babuk, Ryuk et Conti

La méthode d'automatisation d'analyse des malwares exposée ci-contre a été soumise à plusieurs cas d'usages. Les familles de malwares Babuk, Ruyk et Conti ont ainsi été passées au tamis du concept-code et de l'analyse en rétro-ingénierie. Ces virus ont récemment frappé de nombreuses organisations. Issu de la famille des rançongiciels, Babuk est apparu pour la première fois en décembre 2020 avant que son code n'évolue en mai 2021. Observé pour la première fois en 2018, Ruyk a également vu ses fonctionnalités évoluer : en octobre 2020, il aurait été responsable de 75% des attaques sur le secteur américain de la santé [CERT, 2021]. Quant à Conti, il est apparu en 2020 et a, lui aussi, frappé de nombreuses organisations, notamment le système de santé irlandais en mai 2021 (291 victimes revendiquées).

Appliquée à ces virus, la matrice de corrélation par concept-code a démontré que des liens existaient entre plusieurs séries de souches du virus Babuk. Par ailleurs, la technologie a montré qu'il y avait un *persona* de développement identique à Conti sur certains échantillons Ryuk. La chronologie de compilation a notamment permis de visualiser un entrelacement entre Conti et Ruyk.

Cette information est précieuse pour les spécialistes de la *Cyber Threat Intelligence* : en reliant la menace à un échantillon précis, elle implémente la base de données sur les malwares et oriente la stratégie des équipes mobilisées en réponse à incident au sein des organisations.

ou d'un malware déjà ingéré, évitant donc les duplicatas.
Le stockage est pour sa part essentiel. Il doit tout à la fois permettre de stocker des données de type varié, de les récupérer à la demande, mais également de faire évoluer l'infrastructure de stockage de manière agile. À cet effet, le choix a été fait de stocker les métadonnées sur un cluster Elasticsearch, et les données dans un serveur Nexus.

Tagger les données et les transformer en concepts-code

La phase de « taggage » (*tagging*) permet de compléter les informations propres au malware et de les référencer dans Elasticsearch. Deux taggers sont ici mobilisés, métadonnées et règles Yara. Le premier utilisera les métadonnées de la source (tags renseignés sur les plateformes de collecte des virus, notamment sur les familles de malwares) ; quant au second, il utilisera une série de règles destinées à identifier certaines familles.
La phase ultime de transformation en concept-code intervient dès lors que l'ensemble des binaires ont pu être collecté et identifié. Cette étape est celle de la transformation et de la conversion. Elle corrèle les malwares ainsi que leurs familles d'appartenance de manière automatique, via une chaîne de collecte asynchrone permettant de faciliter les reprises, d'interrompre la collecte et de la relancer ultérieurement. Pour chaque source, les recherches de liens sont réalisées toutes les dix minutes, et des tags mis à jour toutes les heures.
Ainsi réalisée, l'automatisation est essentielle. Elle permet actuellement de traiter quelque 8,2 millions de malwares, dont 3,3 millions tagués et 4,4 millions non rejetés. Cette collecte est indubitablement un pas supplémentaire franchi vers la *Cyber Threat Intelligence.*

Le concept-code : une technologie validée scientifiquement

La technologie du concept-code a récemment fait l'objet d'un article scientifique. Co-signé par Frédéric Grelot, Marie Salmon et Sébastien Larinier, cet article a permis de comparer deux méthodes d'analyse des binaires : l'une, manuelle, assurée par un expert de l'école d'ingénieurs en intelligence artificielle EPITA ; l'autre, automatique, par les ingénieurs de la start-up GLIMPS.
Au final, la recherche automatisée et la technologie du concept-code sont validées par l'approche académique manuelle. Menée à petite échelle (une cinquantaine de virus), la comparaison montre que l'automatisation en rétro-ingénierie est non seulement fiable, mais aussi bien plus rapide que l'approche classique. Cette automatisation permet de générer un précieux gain de temps pour la défense des organisations.

Entretien avec Sébastien Larinier

Concept-code « J'ai tout de suite réalisé que les résultats étaient très intéressants »

Concept-code : *« J'ai tout de suite réalisé que les résultats étaient très intéressants »*

Enseignant-chercheur à l'ESIEA, et membre du laboratoire Confiance Numérique et Sécurité, Sébastien Larinier a découvert récemment la technologie du concept-code. Il revient sur les apports de cette innovation reposant sur l'IA, ainsi que sur les objectifs qu'elle permet désormais d'atteindre en termes de cyberdéfense.

Vous travaillez sur l'analyse de virus : dans quel contexte s'inscrivent vos recherches, et sur quoi portent-elles ?

Sébastien Larinier : J'enseigne l'analyse de virus au sein de l'ESIEA, école d'ingénieur d'un numérique utile. Il s'agit de l'une des 204 écoles d'ingénieurs françaises accréditées à délivrer un diplôme d'ingénieur, et nous sommes implantés sur Paris, Ivry-Sur-Seine et Laval. J'y travaille en tant qu'enseignant et chercheur, au sein du laboratoire CNS (Confiance Numérique et Sécurité). Mon métier est centré sur les virus et les ransomwares, et tout particulièrement sur les attaques dites d'Etat, que l'on appelle également les APT. La Chine et l'Inde font partie des pays sur lesquels j'effectue mes recherches.

Comment avez-vous été amené à découvrir le concept-code ?

S.L : Tout à fait par hasard. J'ai été interpellé sur Twitter par Frédéric Grelot, l'un des ingénieurs à la base de cette technologie. Celui-ci avait certainement identifié mon cœur de recherches et m'a invité à prendre connaissance de la matrice de corrélation portant sur Babuk, l'un des virus sur lesquels je travaille et que je connais le mieux. En comparant les notes que j'avais prises avec les résultats de cette matrice, j'ai tout de suite réalisé que les résultats obtenus étaient vraiment intéressants.

Pour quelles raisons ?

S.L : Le cas du virus Babuk est en réalité un cas d'école sur lequel je travaille un peu comme un généticien pourrait le faire avec des virus humains : en recherchant des modifications de séquences. Cela me permet de faire ce que l'on appelle du suivi de code et de voir les différentes versions de ce virus. En génétique, on appellerait cela de la phylogénie… L'outil lié au concept-code a ceci de précieux qu'il permet d'identifier toutes les familles de variants de manière automatique, notamment via les statistiques générées. Pour un chercheur, c'est un gain de temps. On peut désormais passer les familles de virus dans une sorte de moulinette du concept-code et obtenir automatiquement l'ensemble des variants si ces derniers ne subissent pas de modifications par des programmes appelés « packer ». De plus, l'outil a permis de mettre en évidence, sur la famille de Babuk, les échantillons qui faisaient le lien entre différentes versions.

La plus-value du concept-code résulte donc dans l'automatisation ?

S.L : En effet, mais pas seulement. Certes, la dimension automatique est un élément majeur de cette technologie. La vitesse, l'efficacité et la sûreté également. Imaginez : jusqu'alors, je réalisais mes calculs de manière artisanale. C'est long, fastidieux, et la fatigue générée chez moi peut entraîner des erreurs. Les concepts-code reproduisent ainsi de manière rapide ce travail et cela sur des millions de virus. De plus, le concept-code permet de pointer des éléments intéressants et des pivots. Après, bien sûr, il faut que l'analyste vérifie par cluster ce qui rend communes toutes les familles de virus collectées. Mais au moins, une fois cette vérification faite, cela permet de travailler sur un seul « individu », c'est-à-dire sur un seul malware. On va dès lors pouvoir émettre des hypothèses, et celles-ci vont se répercuter sur l'ensemble de ces individus. En tant qu'analyste, on aime bien cela : gain de temps, efficacité, sûreté… L'algorithme ne se trompera jamais. Encore faut-il que ledit algorithme ait été bien conçu. C'est bien ce qui se passe avec la technologie du concept-code sur les cas sur lesquels nous avons pu travailler, j'ai pu m'en rendre compte.

Au final, que permet concrètement cette technologie ?

S.L : Les concepts-code ont pu rendre des modèles d'intelligence artificielle efficients sur l'analyse de virus informatiques et leurs similarités. C'était un problème devant lequel nous nous trouvions depuis quelques années, mais jusqu'alors nous n'utilisions pas l'IA ou alors de manière extrêmement complexe, en dépassant rarement la porte des laboratoires. Le concept-code nous dit également d'autres choses, de nature intra-familiale celles-là. En

fait, il faut bien comprendre que les cyberattaquants utilisent des techniques de tromperie en modifiant la partie génétique du code. Ils usent de ruses qui leur sont propres. En tant qu'outil, le concept-code arrive à démontrer que des stratégies identiques sont déployées à l'intérieur de familles de malwares différentes. De ce fait, nous pouvons montrer que deux familles de virus peuvent être le fait d'un seul et même groupe de cyberattaquants. Ça pointe du doigt les techniques qu'ils utilisent, et cela les démasque. Une fois qu'on a compris cela, on peut organiser une défense efficace : générer par exemple une signature d'antivirus. On bloque donc le malware plus tôt, et on le fait massivement lorsque des familles entières utilisent un procédé identique. C'est également valable pour les sondes de détection d'intrusion et les EDR (Endpoint Detection Response).

Détection de malwares : une course de vitesse

En matière de cybersécurité, la vitesse est un atout majeur. Face à une attaque, il est en effet impératif de réagir au plus vite, d'abord et avant tout au niveau de la détection. « *Actuellement, on observe que les algorithmes peuvent être efficaces sur la détection mais cela prend du temps* », explique Valérian Comiti, Ingénieur R&D, Directeur des Opérations au sein de la start-up GLIMPS. « *Quand une solution de défense a un doute sur un dossier, elle va émuler le code afin de tenter de déterminer ce qui est en train de se passer. Cette opération peut être plus ou moins longue, entre une minute et une heure afin d'exécuter le binaire. C'est très souvent suffisant pour que le malware se propage dans les systèmes et fasse des dégâts importants.* » Face à cela, le concept-code dispose d'un atout : il « lit » le code de manière linéaire et en dégage l'ensemble des concepts dont il est porteur, sans exclusion aucune. « L*a vitesse d'extraction des concepts est très rapide, de l'ordre de trois secondes maximum* », poursuit Valérian Comiti. Pour l'exprimer autrement, la technologie du concept-code fait gagner un temps précieux en se concentrant exclusivement sur l'histoire dont est porteur le virus, et non sur l'ensemble des mots qu'il emploie ni sur sa grammaire. « *C'est la différence qu'il y a entre l'émulation du code et sa lecture. Si on devait émuler et simuler l'exécution intégralement cela prendrait du temps. En conceptualisant le code dès sa lecture, nous opérons une action synthétique et atteignons des vitesses de détection sans commune mesure avec ce qui se fait actuellement.* »

Demain : comment se défendre face aux malwares ?

Le variant, danger pour l'ensemble des organisations

Les variants constituent la véritable problématique actuellement observée en matière de cybersécurité. Parmi les quelque 2240 attaques recensées quotidiennement (soit une attaque en moyenne toutes les 39 secondes), on estime à 99,9% le taux de malwares qui constituent de simples variants de virus natifs. « *La production de variants a ceci de spécifique qu'elle permet de passer les barrières de sécurité traditionnelles* », expliquent Frédéric Grelot et Cyrille Vignon, Ingénieurs spécialisés en cybersécurité. Selon certains analystes, chaque jour, 230 000 variants de malwares seraient ainsi produits, soit une évolution quantitative massive.

Le fait que la grammaire de ces virus ait été légèrement modifiée est à elle seule un gage de « succès » face aux moyens de défense classiques. Un antivirus est en effet programmé pour repérer à la fois une « langue » et une grammaire, et non pour identifier le code en tant qu'histoire… « *En fait, toutes les attaques qui réussissent actuellement sont celles qui utilisent des variants : c'est l'essence même de leur succès* », indiquent Frédéric Grelot et Cyrille Vignon. D'où la nécessité d'aller débusquer le mal au cœur même de la conception du code, c'est-à-dire de la narration du malware.

Passer l'intégralité des variants au tamis de la technologie du concept-code est aujourd'hui l'une des seules techniques efficaces. Il est certain que la production de malwares est encore appelée à se développer dans les mois à venir, via des cybercriminels aux process de plus en plus structurés, pour ne pas dire industrialisés.

Quel est le système de défense le plus approprié pour faire face aux cyber-attaques dont les organisations sont actuellement l'objet ? Nous avons interrogé deux ingénieurs spécialisés sur le sujet, en lien avec la technologie du concept-code ainsi qu'à son évolution vers la notion de généalogie binaire.

Profusion des attaques, multiplicité des failles, diversité des technologies engagées… Existe-t-il, face à l'essor sans précédent des malwares, une défense optimale ? À en croire Valérian Comiti, ingénieur Recherche & Développement directeur des opérations de GLIMPS, la stratégie de défense est aussi complexe et nuancée à mettre en œuvre que les attaques le sont elles-mêmes.

Multiplier les solutions de sécurité

« *La défense la plus optimale sera celle qui multipliera en les intégrant ensemble les solutions de sécurité* », explique le spécialiste. « *Il faut mettre en place une grande variété de technologies de défense complémentaires les unes avec les autres.* » Les options sont donc plurielles : observation de l'activité des postes grâces aux EDR, analyse des flux entrants comme les flux sortants, concentration sur les comportements, analyse fine de ce qu'il se passe à l'intérieur même des flux… « *Certaines solutions excellent à détecter les logiciels*

qui sont strictement malveillants ; d'autres permettent d'identifier les variants, comme celle du concept-code sur laquelle je travaille actuellement. Encore une fois, il n'y a pas une seule solution : la solution est dans la bonne articulation de lignes de défense variées. » Une certitude cependant : la technologie est un élément clé de la défense qui se met actuellement en place au sein de nombreuses organisations privées et publiques. « *Grâce au concept-code, nous disposons d'un outil différenciant qui va poser d'importants problèmes aux cybercriminels* », souligne Valérian Comiti. « *C'est un coup d'échec et mat en ce sens où la technologie que nous employons lit le code et définit de manière immédiate si le système est infecté par un variant de malware.* » Difficile pour les attaquants de contourner les organisations qui utilisent le concept-code : un binaire ne peut passer inaperçu auprès d'un système armé d'Intelligence Artificielle. À ce stade de la lutte que mènent certaines start-up – notamment françaises – face aux cyber-attaques, le concept-code offre un avantage indéniable.

La généalogie binaire, une nouvelle frontière

En sera-t-il de même demain ? Sur ce point, Frédéric Grelot se veut résolument optimiste. « *Nous voyons d'ores et déjà que deux types d'attaques sont aujourd'hui possibles* », explique le polytechnicien, ingénieur en physique et en informatique. « *Le premier cas d'attaque est celui, classique pour nous, où le cyberattaquant a modifié un malware déjà existant. Dans ce cas, le concept-code va démasquer le variant. Cela concerne actuellement la très large majorité des attaques qui se produisent.* » Mais que se passe-t-il dans le cas où le malware est d'un type nouveau, doté d'un code qui n'a jamais été

répertorié ? « *C'est ici qu'entre en jeu la complémentarité avec le système de défense, un EDR par exemple qui va identifier un comportement malveillant sur un poste isolée. Le concept-code nouveau qui sera alors généré permettra de passer tout le reste du système d'information au tamis de la vérification : nous déterminerons ainsi si d'autres machines sont infectées, et neutraliserons le malware.* » Peu à peu se tissent ainsi des liens entre les différents types de logiciels malveillants. Ascendance, descendance, lignée… L'évolution des travaux qui sont actuellement menés par les ingénieurs R&D spécialisés en cyber-sécurité renvoient très directement à la notion de filiation entre malwares, pour ne pas dire à la généalogie binaire. En l'espèce, le parallèle avec la généalogie humaine ou animale peut être fait : l'enjeu consiste à définir des liens de parenté entre telle ou telle famille, sur la base du fait que le génome est en capacité d'être décodé. « *Il est tout à fait possible de faire apparaître des liens entre des concepts-code différents afin de voir s'ils ont des liens de parenté les uns avec les autres* », conclut Frédéric Grelot. « *C'est ce que nous envisageons de faire dans le futur. Même s'il n'y a pas encore d'urgence à le faire pour l'instant, nous savons qu'il s'agit là d'une piste à explorer si nous voulons franchir un palier supplémentaire dans nos recherches.* »

Comment le concept code s'adapte à l'environnement Sécurité des Systèmes d'Information (SSI) des organisations

« *Nous travaillons au cas par cas, sans process type car tous les environnements sont différents… Nous faisons face à des problématiques, que nous résolvons.* » Ingénieur en cybersécurité, Chief Technical Officer au sein d'une start-up spécialisée dans le concept-code, Jérémy Bouétard l'affirme : même avec une diversité d'environnements, il n'est pas très difficile d'intégrer cette nouvelle technologie au sein d'un système d'information. « *La solution est relativement simple à interconnecter dans la mesure où nos équipes ont automatisé et documenté toutes les étapes… Le plus difficile c'est sans doute que le logiciel de l'entreprise soit ouvert. Après, il s'agit de lancer un script qui va demander le nom de domaine et configurer le serveur.* » Une fois l'opération réalisée, l'organisation dispose d'un tableau de bord (dashboard) qui lui permet en trois secondes de savoir si un ou plusieurs malwares sont présents. Certaines structures souhaitent aller plus loin en prenant connaissance des zones au sein desquelles réside le binaire malveillant…

« *Dans certains cas, lorsque les organisations sont importantes, il va falloir coordonner l'ensemble de ces services et directions. Mais je dois dire que cela ne coince jamais techniquement.* »

GLOSSAIRE

Concept-code :

La conceptualisation du code est une technologie récente consistant à identifier de manière automatisée, au sein d'un malware, la narration de celui-ci. Après avoir été désassemblé grâce au *Reverse Engineering*, le code livre ainsi l'histoire dont il est le détenteur, par-delà les chiffres qui le composent et qui en sont tout à la fois les mots et leur grammaire. Ainsi, si le malware était un livre racontant un conte, la conceptualisation du code en constituerait l'histoire. Ce procédé permet de repérer des familles de malwares dans la mesure où ces logiciels malveillants sont à 99% le produit de variants (variations) d'une histoire déjà connue. Le concept-code établit ainsi qu'une histoire est identique à une autre à 5%, 10% ou 15%, permettant de stopper net la progression d'un malware. À noter que la conceptualisation du code était ces dernières années déjà expérimentée dans le champ académique, mais de manière manuelle. L'élément nouveau vient de l'automatisation de cette technologie, qui s'appuie tout à la fois sur le *Reverse Engineering*, l'Intelligence Artificielle et le *Machine Learning*.

Intelligence Artificielle :

L'IA est un ensemble de technologie différentes qui fonctionnent ensemble dans le but de permettre aux machines de percevoir, comprendre, agir et apprendre à des niveaux d'intelligence tendant à celle des humains. Plusieurs sous-ensembles technologiques font partie de l'IA, tels que le traitement automatique du langage naturel (NLP, *Natural Language Processing*) ou le *Machine Learning* (ML). L'Intelligence Artificielle peut être soit étroite (pour des applications concrètes tels qu'un assistant virtuel) soit générale (c'est-à-dire « forte », capable de réflexions stratégiques, abstraites et créatives dans un ensemble de tâches complexes). Dans le cas de la détection de malwares par la technologie du concept code, l'IA joue un rôle central en compilant l'ensemble des familles de logiciels malveillants sur une dizaine d'années. Cela représente plusieurs millions de données.

Machine Learning :

Le *Machine Learning* constitue l'une des formes de l'Intelligence Artificielle, une sous-catégorie de celle-ci visant à automatiser le processus de création de modèles analytiques. Le *Machine Learning* permet aux machines de s'adapter à de nouveaux scénarii de manière autonome. Il permet notamment une gestion intelligente de données massives (Big Data), ce qui dans le cadre de la recherche de malwares est un atout précieux. À la base, le *Machine Learning* prend la forme d'un ordinateur qui examine des données et identifie des schémas avant d'accomplir la tâche pour laquelle il a été défini (en l'occurrence ici la détection de logiciels malveillants).

Reverse Engineering :

Le *Reverse Engineering* (ou rétro-ingénierie) est la technique grâce à laquelle un système peut être analysé en sens opposé à sa réalisation, pour déduire les méthodes et techniques qui le constituent en partant du produit final. Dans le champ des systèmes informatiques, cette étude et cette analyse peuvent être appliquées aux logiciels, qu'ils soient malveillants (malwares) ou non (goodwares). Le Reverse Engineering logiciel consiste ainsi à inverser le code machine d'un malware afin de retrouver le niveau de compréhension du code source dans lequel il a été écrit, en utilisant des instructions en langage de programmation. La rétro-ingénierie des logiciels peut être utilisée pour comprendre un programme parce que son code source a été perdu, ou parce qu'il n'est pas connu. Ce faisant, le *Reverse Engineering* permet d'identifier le contenu malveillant d'un programme.

Matrice de corrélation (par concept-code) :

Matrice à deux dimensions regroupant toutes les valeurs de similarité calculées pour un ensemble de binaires, en comparant leurs concepts-code deux à deux. Cette matrice fournit une représentation concise des similarités qui sert de base à l'analyse de liens entre les malwares d'une même famille ou de familles différentes.

RÉFÉRENCES BIBLIOGRAPHIQUES

BLOOMBERG, *« Hackers Bearing Down on U.S. Hospitals Have More Attacks Planned »*, 30 octobre 2020. URL : https://www.bloomberg.com/news/articles/2020-10-30/hackers-bearing-down-on-u-s-hospitals-have-more-attacks-planned

CERT, *« Etat de la menace rançongiciel à l'encontre des entreprises et des institutions »*, sept. 2021. URL : https://www.cert.ssi.gouv.fr/uploads/CERTFR-2021-CTI-001.pdf

CERT, *« Le Groupe Cybercriminel TA505 », 10 février 2021 (version actualisée).* REDMINE : https://www.cert.ssi.gouv.fr/cti/CERTFR-2020-CTI-006/

CERT, *« Etat de la menace rançongiciels à l'encontre des entreprises et institutions »*, 6 octobre 2021. URL : https://www.cert.ssi.gouv.fr/cti/CERTFR-2021-CTI-001/

DIGITAL SHADOWS, *« DarkSide : The New Ransomware Group behind Highly Targeted Attacks »*. 22 septembre 2020. URL : https://www.digitalshadows.com/blog-and-research/darkside-the-new-ransomware-group-behind-highly-targeted-attacks/

ENISA, *« ENISA Threat Landscape 2020 – Ransomware »*, 23 octobre 2020. URL : https://www.enisa.europa.eu/publications/ransomware.

FINCEN, *« Financial Trend Analysis. Ransomware Trends in Bank Secrecy Act Data Between January 2021 and June 2021 »*, octobre 2021. URL : https://www.fincen.gov/sites/default/files/2021-10/Financial%20Trend%20Analysis_Ransomware%20508%20FINAL.pdf

FRANCE INFO, *« Cyberattaques : le nombre de piratages a quadruplé l'année dernière »*, 16 février 2021. URL : https://www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/cyberattaques-le-nombre-de-piratages-a-quadruple-l-annee-derniere-selon-un-expert-en-cybersecurite_4299525.html

Gesny Olivier, *« Capter l'IA de demain au regard des enjeux de cyberdéfense »*, Revue Défense Nationale, 2019/5 (N° 820), p. 38-42. DOI : 10.3917/rdna.820.0038. URL : https://www.cairn.info/revue-defense-nationale-2019-5-page-38.htm

GRELOT Frédéric, LARINIER Sébastien et SALMON Marie, *« Automatisation de l'analyse des binaires : de la collecte source ouvert à la Threat Intelligence »*, 16 novembre 2021, salon C&AESAR 2021. URL : https://conf.researchr.org/details/cesar-2021/call-for-papers/14/Automatisation-de-l-analyse-de-binaires-de-la-collecte-source-ouverte-la-Threat-I

INSTITUT MONTAIGNE, *« Cybercrime : plongée dans l'écosystème »*, 2021. URL : https://www.institutmontaigne.org/blog/cybercrime-plongee-dans-lecosysteme

Informatique News, *« Tableau de bord pour visualiser les cyberattaques à travers le monde »*, 27 août 2021. URL : https://www.informatiquenews.fr/15-tableaux-de-bord-pour-visualiser-les-cyberattaques-a-travers-le-monde-72060

PACKETLABS, *« Qu'est-ce que la chasse au gros gibier ? »*, 2012. URL : https://www.packetlabs.net/big-game-hunting/

REYNAUD Florian, *« Qui sont les hackers qui ont récemment paralysé une partie du système de santé de l'Irlande avec un rançongiciel ? »*, Le Monde, 15 mai 2021. URL : https://www.lemonde.fr/pixels/article/2021/05/15/qui-sont-les-pirates-qui-ont-frappe-le-systeme-de-sante-irlandais-avec-un-rancongiciel_6080311_4408996.html

SECURE LIST, *« Targeted Ransomware : It's Not Just about Encrypting Your Data ! »*, 11 novembre 2020. URL : https://securelist.com/targeted-ransomware-encrypting-data/99255/

THREATPOST, *« Egregor Ransomware Threatens 'Mass-Media' Release of Corporate Data »*, 2 octobre 2020. URL : https://threatpost.com/egregor-ransomware-mass-media-corporate-data/159816/.

TWITTER. @malwrhunerteam. 8 janvier 2021. URL : https://twitter.com/malwrhunterteam/status/1347458694053822464.

VADESECURE, *« Ransomware as a service (RaaS) : une activité illicite qui a désormais pignon sur rue »*, 19 mars 2020. URL : https://www.vadesecure.com/fr/blog/ransomware-as-a-service-raas-une-activite-illicite-qui-a-desormais-pignon-sur-rue

Web : https://www.glimps.fr/
Linkedin : https://www.linkedin.com/company/glimpsre/
Twitter : https://twitter.com/GlimpsRe
Email : contact@glimps.re